American Economic Review 2012, 102(7): 3628–3651
http://dx.doi.org/10.1257/aer.102.7.3628

The Effect of Evaluation on Teacher Performance[†]

By Eric S. Taylor and John H. Tyler*

The effect of evaluation on employee performance has been a long-standing interest shared by researchers, firms, and policymakers across sectors. Still, relatively little empirical attention has been given to the potential long-run effects of performance evaluations including employee skill development. This topic is increasingly salient for American public schools as over the past decade evaluating teacher effectiveness has become a dominant theme in the education sector. The emphasis on evaluation is motivated by two oft-paired empirical conclusions: (i) teachers vary greatly in their ability to promote student achievement growth, but (ii) observable teacher characteristics like graduate education and experience (beyond the first few years) are not typically correlated with increased productivity. Many researchers and policymakers have suggested that, under these conditions, the only way to adjust the teacher distribution for the better is to gather information on individual productivity through evaluation and then dismiss low performers.

This paper offers evidence that evaluation can shift the teacher effectiveness distribution through a different mechanism: by improving teacher skill, effort, or both in ways that persist long-run. We study a sample of mid-career math teachers in the Cincinnati Public Schools (CPS) who were assigned to evaluation in a manner that permits a quasi-experimental analysis. All teachers in our sample were evaluated by a year-long classroom observation–based program, the treatment, between 2003–2004 and 2009–2010; the timing of each teacher's specific evaluation year was determined years earlier by a district planning process. To this setting we add measures of student achievement, which were not part of the evaluation, and use the within-teacher over-time variation to compare teacher performance before, during, and after their evaluation year.

We find that teachers are more productive during the school year when they are being evaluated, but even more productive in the years after evaluation. A student taught by a teacher after that teacher has been through the Cincinnati evaluation will score about 10 percent of a standard deviation higher in math than a similar student taught by the same teacher before the teacher was evaluated.

Under our identification strategy, these estimates may be biased by patterns of student assignment that favor previously evaluated teachers, or by preexisting positive

*Taylor: Stanford University, 520 Galvez Mall, CERAS Building, Room 509, Stanford, CA 94305 (e-mail: erictaylor@stanford.edu); Tyler: Brown University, Box 1938, Providence, RI 02905 (e-mail: john_tyler@brown.edu). The authors would like to thank Eric Bettinger, Ken Chay, David Figlio, Caroline Hoxby, Susan Moore Johnson, Susanna Loeb, Doug Staiger, two anonymous reviewers, and seminar participants at Wellesley, Stanford, and the NBER Education Program for helpful comments on previous drafts of this paper. The research reported here was supported in part by the Institute of Education Sciences, US Department of Education, through Grant R305C090023 to the President and Fellows of Harvard College. The opinions expressed are those of the authors and do not represent views of the Institute or the US Department of Education. We also gratefully acknowledge the Center for Education Policy Research at Harvard University, the Joyce Foundation for their generous support of this project, and the cooperation and support of the Cincinnati Public Schools.

[†] To view additional materials, visit the article page at http://dx.doi.org/10.1257/aer.102.7.3628.

trends in teacher performance. We investigate these threats through event studies and comparisons of observable teacher and student characteristics across treatment groups, and find little evidence of bias.

While the data do not provide information that allows us to identify the exact mechanisms driving the results, these gains in teacher productivity are consistent with a model whereby teachers learn new information about their own performance during the evaluation and subsequently develop new skills, or increase long-run effort, or both. This information mechanism suggests that the general pattern of results may extend to other sectors and professions when individualized performance information is scarce.

The teachers in our sample—who were in the middle of their careers and had not been evaluated systematically for some years—may have been particularly responsive to the influx of new personalized performance information created by classroom observation–based evaluation. Effects may be smaller where personalized evaluative feedback is more regular.

Nevertheless, the results of this analysis contrast sharply with the widely held perspective that the effectiveness of individual teachers cannot be changed much after the first few years on the job, suggesting a role for teacher evaluation beyond selective retention. Indeed, our estimates indicate that postevaluation improvements in performance were largest for teachers whose performance was weakest prior to evaluation, suggesting that teacher evaluation may be an effective professional development tool.

I. Related Literature and Mechanisms

Motivated by large differences in productivity from teacher to teacher (see Hanushek and Rivkin 2010 for a review),[1] research efforts have tried to identify predictors of teacher productivity that could be used to inform human resource decisions. Many of the intuitive candidates, like the possession of a graduate degree or teacher professional development, have proven to be dead ends, and while teachers do improve with experience, the returns to experience appear to level off relatively quickly (Hanushek 1986, 1997; Rockoff 2004; Jacob 2007; Yoon et al. 2007; Rockoff et al. 2011).

Absent evidence that information traditionally found in a teacher's personnel file can predict effectiveness, recent research efforts have turned to measuring individual teacher performance more directly. The literature suggests that various measures of individual teacher performance are promising sources of information for human resource decisions. Some argue that the most direct and objective evidence of teacher performance are so called "value-added" measures based on student test score gains. Using student test score–based measures, while intuitive, is not always possible and not without research and political controversy (Glazerman et al. 2010). Encouragingly, however, several other performance appraisal approaches appear to be good predictors of a teacher's ability to promote student achievement. These include subjective ratings by principals and other experienced educators who are familiar with the teacher's day-to-day work (Jacob and Lefgren 2008; Rockoff and Speroni 2010; Rockoff et al.

[1] While estimates across researchers and settings are relatively consistent, there remain questions about the empirical identification (Rothstein 2010, Todd and Wolpin 2003).

2012), ratings based on structured classroom observation (Grossman et al. 2010; Kane et al. 2011), student surveys (Kane and Cantrell 2010), and assessments of teachers by external evaluators like the National Board for Professional Teaching Standards (Goldhaber and Anthony 2007; Cantrell et al. 2008). On the other hand, the formal status quo teacher evaluation programs currently utilized by most districts are perfunctory at best and conceal the variation in performance (Weisberg et al. 2009).

Assuming that improved teacher performance measures can be adopted, much of the discussion regarding what to do with those measures has focused on selective dismissal. Predictions of the net gain of selective dismissal and retention are mixed (Gordon, Kane, and Staiger 2006; Goldhaber and Hansen 2010; Hanushek 2011) and empirical evidence is very rare. One exception is a recent experimental study by Rockoff et al. (2012) where principals in the treatment group were given objective student test score–based ratings of teachers. A first result of the study is that these principals systematically adjusted their subjective assessments of teachers to more closely match the objective information. Subsequently, treatment schools in the study experienced greater turnover of low-performing teachers and made small gains in math achievement relative to the control schools. These results are consistent with a model where principals improve their instructional staff via selective retention that is based on performance data. On the other hand, Staiger and Rockoff (2010) describe weak assumptions under which the optimal turnover of novice teachers would be quite high.

Other discussions propose tying evaluation scores to incentive pay, and various districts, including, notably, Denver and Houston, have instituted merit pay plans that do this. While theoretically promising, the early work in this area suggests mixed results (Springer et al. 2010; Neal 2011).

The broader personnel economics literature suggests other mechanisms, beyond selection and monetary incentives, through which employee evaluation might lead to productivity gains. There are, however, competing views in this literature on how evaluation optimally achieves such gains. One perspective, motivated by the traditional principal-agent framework, holds that evaluation should closely link performance with rewards and punishments in a way that directly incentivizes employee effort. From this perspective, productivity effects are expected to be proximate to the period and content of the evaluation; mechanisms for lasting productivity gains are not generally addressed in these models.[2]

An alternative perspective focuses on using performance appraisal as an integral part of long-run employee development rather than as a tool in a rewards-and-punishment incentive scheme. This human resource development view of evaluation posits that evaluation linked to rewards and punishment can subvert the developmental aspects of appraisal because the employee being incentivized by rewards and punishment–related evaluation views the process as judgmental and punitive (Armstrong 2000). Since current evaluation programs rarely lead to rewards or punishments for teachers (Weisberg et al. 2009), evaluation in the education sector may be better understood from the developmental perspective rather than a traditional principal-agent model.

[2] An example of a study of the proximate effects of subjective evaluation on worker input that could be expected to impact productivity from the human resource management literature is by Engellandt and Riphahn (2011). They found that employees in one international company respond to incentive mechanisms in subjective supervisor evaluations by supplying more effort to the job, but it remains unclear how those changes in effort affected output or whether the extra effort was nontransient.

It is also the case that teacher performance may be particularly susceptible to the developmental aspects of evaluation. Dixit (2002) posits that teachers are generally "motivated agents," and to the extent that this is true we would expect teachers to act on information that could improve individual performance. Yet, individualized, specific information about one's performance seems especially scarce in the teaching profession (Weisberg et al. 2009), suggesting that a lack of information on *how* to improve could be a substantial barrier to individual productivity gains among teachers. Well-designed evaluation might provide new information to fill that knowledge gap in several ways. First, teachers could gain information through the formal scoring and feedback routines of an evaluation program. Second, evaluation could encourage teachers to be generally more self-reflective regardless of the evaluative criteria. Third, the evaluation process could create more opportunities for conversations with other teachers and administrators about effective practices. Additionally, programs that use multiple evaluators including peers, as is the case in Cincinnati, may result in both more accurate appraisals and more take-up by the individuals evaluated (Kluger and DeNisi 1996; Kimball 2002).

To improve performance, however, the information from evaluation must be correct (in the sense that the changes implied will improve effectiveness if acted on). As mentioned above, a small but growing number of empirical studies have found meaningful correlations between observed teacher practices, as measured by evaluative criteria, and student achievement growth. This includes the Cincinnati program that is the setting of this paper. Kane et al. (2011) found that teachers who received higher classroom practice scores on Cincinnati's evaluation rubric also systematically had higher test-score value-added. Student math achievement was 0.087 standard deviations higher for teachers' whose overall evaluation score was one standard deviation higher (the effect for reading was 0.78).[3,4] This cross-sectional relationship suggests that Cincinnati's evaluation program scores teachers and provides feedback on teaching skills that are associated with promoting higher student achievement. To the extent that teachers improve in those skills, we would anticipate improvements in performance as measured by value-added to student achievement. While the Kane et al. (2011) study documented a relationship between evaluation scores and value-added extant at the time of evaluation, this paper asks and tests a separate question: does the process of going through a year-long evaluation cycle improve teacher effectiveness, as measured by value-added?

In addition to new, individualized information provided to teachers, there may be other mechanisms through which evaluation could impact teacher effectiveness. The process of defining and communicating the evaluative criteria to employees may result in a greater focus on (presumably correct) practices among the teachers of a school or district (Milanowski and Heneman 2001). Alternatively, teachers may increase their effort level during evaluation as a response to traditional incentives only to find that a higher level is a preferable long-run equilibrium.

These mechanisms for lasting improvements do not preclude teacher responses to the proximate incentives of evaluation. In practice, however, the formal stakes

[3] The mean overall evaluation score was 3.21 out of 4 with a standard deviation of 0.433.

[4] Holtzapple (2003), Milanowski (2004), and Milanowski, Kimball, and White (2004) demonstrated a positive relationship between formal evaluation scores in the Cincinnati system and achievement early in the program's life.

of Cincinnati's evaluation program are relatively weak, as we discuss in the next section. Additionally, while individual teachers and their evaluators gain much new information in the evaluation process, administrators with the authority to reward or punish teachers based on evaluation results learn a teacher's final overall scores in only four areas, and these final scores have little meaningful variation. While we cannot rule out a proximate response, in the end our results are more consistent with skill development and long-run change.

As the discussion to this point suggests, there is reason to expect that well-designed teacher evaluation programs could have a direct and lasting effect on individual teacher performance. To our knowledge, this study is the first to test this hypothesis empirically. Given the limitations of the data at hand, we can only speculate on the mechanisms through which evaluation might impact subsequent performance. The setting suggests, however, that the individualized performance feedback experienced teachers receive in the evaluation process is a likely mechanism. Regardless of the mechanism, however, the results of this study highlight returns to teacher evaluation outside the more-discussed mechanisms of monetary incentives and selective dismissal.

II. Data and Setting

The data for our analysis come from the Cincinnati Public Schools. In the 2000–2001 school year, Cincinnati launched the Teacher Evaluation System (TES) in which teachers' performance in and out of the classroom is evaluated through classroom observations and a review of work products. During a year-long process, each teacher is evaluated by a school administrator and a peer teacher. Owing mostly to cost, however, each teacher is typically evaluated only every five years.

During the TES evaluation year, teachers are typically observed in the classroom and scored four times: three times by an assigned peer evaluator—high-performing, experienced teachers who are external to the school—and once by the principal or another school administrator. Teachers are informed of the week during which the first observation will occur, with all other observations being unannounced. The evaluation measures dozens of specific skills and practices covering classroom management, instruction, content knowledge, and planning, among other topics. Evaluators use a scoring rubric, based on Charlotte Danielson's *Enhancing Professional Practice: A Framework for Teaching* (1996), which describes performance of each skill and practice at four levels: "Distinguished," "Proficient," "Basic," and "Unsatisfactory." For example, standard 3.4.B addresses the use of questions in instructional settings:

- Distinguished: "Teacher routinely asks thought-provoking questions at the evaluative, synthesis, and/or analysis levels that focus on the objectives of the lesson. Teacher seeks clarification and elaboration through additional questions. Teacher provides appropriate wait time."
- Proficient: "Teacher asks thought-provoking questions at the evaluative, synthesis, and/or analysis levels that focus on the objectives of the lesson. Teacher seeks clarification through additional questions. Teacher provides appropriate wait time."

- Basic: "Teacher asks questions that are relevant to the objectives of the lesson. Teacher asks follow-up questions. Teacher is inconsistent in providing appropriate wait time."
- Unsatisfactory: "Teacher frequently asks questions that are inappropriate to objectives of the lesson. Teacher frequently does not ask follow-up questions. Teacher answers own questions. Teacher frequently does not provide appropriate wait time." [5]

Both the peer evaluators and administrators complete an intensive TES evaluator training course, and must accurately score videotaped teaching examples to check interrater reliability.

After each classroom observation, peer evaluators and administrators provide written feedback to the teacher, and meet with the teacher at least once to discuss the results. At the end of the evaluation school year a final summative score in each of four domains of practice is calculated and presented to the evaluated teacher.[6] Only these final scores carry explicit consequences. For beginning teachers (those evaluated in their first and their fourth years), a poor evaluation could result in nonrenewal of their contract, while a successful evaluation is required before receiving tenure. For tenured teachers, evaluation scores determine eligibility for some promotions or additional tenure protection, or, in the case of very low scores, placement in the peer assistance program with a small risk of termination.

Despite the training and detailed rubric provided to evaluators, the TES program nevertheless experiences some of the leniency bias typical of many other subjective evaluation programs generally (Prendergast 1999) and teacher evaluations particularly (Weisberg et al. 2009). More than 90 percent of teachers receive final overall TES scores in the "Distinguished" or "Proficient" categories. Leniency is much less frequent in the individual rubric items and individual observations. We hypothesize that this micro-level evaluation feedback is more important to lasting performance improvements.

The description of Cincinnati's program may, to some, seem more structured than is suggested by the term "subjective evaluation." Nevertheless, TES is more appropriately studied as a subjective, rather than an objective, evaluation. First, the evaluation is designed to measure performance on dimensions that require informed observation of behavior in context: dimensions that do not yield to standardized measures as do things like widgets produced, sales revenue, or student test score gains. Second, the evaluators' judgments and associated scores cannot, strictly speaking, be verified by an outside party even if they may be more reliable than judgments unguided by rubrics and training. Third, the evaluation is designed to measure performance on inputs to production, not outcomes.

As mentioned above, teachers only undergo comprehensive evaluation periodically.[7] Every teacher newly hired by the district, regardless of experience, is evaluated

[5] The complete TES rubric is available on the Cincinnati Public Schools website: http://www.cps-k12.org/employment/tchreval/stndsrubrics.pdf.

[6] For more details on this final scoring process, see Kane et al. (2011).

[7] In years when teachers are not undergoing a full TES evaluation they do receive an annual evaluation from a school administrator. These annual evaluations are more typical of teacher evaluation in other school districts (Weisberg et al. 2009). The annual evaluations are essentially perfunctory with nearly all teachers receiving a "pass-

Table 1—First Scheduled TES Participation Year for Veteran Teachers

Year of hire	Scheduled evaluation year	Anticipated experience at time of evaluation[a]
1999–2000	2006–2007	8 years
1998–1999	2007–2008	10 years
1997–1998	2005–2006	9 years
1996–1997	2006–2007	11 years
1995–1996	2007–2008	13 years
1994–1995	2008–2009	15 years
1993–1994	2009–2010	17 years

[a] Teachers who take a leave of absence, or began employment at CPS with prior experience, would have different levels of experience.

during their first year working in Cincinnati schools. Teachers are also evaluated just prior to receiving tenure, typically their fourth year after being hired, and every fifth year after achieving tenure. Teachers hired before the TES program began in 2000–2001 were not first evaluated until some years into the life of program. These phased-in teachers form our analysis sample.

A. *Analysis Sample*

Our analysis spans the 2003–2004 through 2009–2010 school years and our sample is composed of fourth through eighth grade math teachers (and their students) who were hired by Cincinnati public schools between 1993–1994 and 1999–2000. We limit our analysis to this sample of mid-career math teachers for three reasons, each bearing on identification. First, for teachers hired before the new TES program began in 2000–2001, the timing of their first TES evaluation was determined largely by a "phase-in" schedule, detailed in Table 1. This schedule, determined during the TES program's planning stages, set the year of first evaluation based on a teacher's year of hire, thus reducing the potential for bias that would arise if the timing of evaluation coincided with a favorable class assignment.[8] Second, as Table 1 shows, the timing of evaluation was determined by year of hire, not experience level, in a pattern such that teachers in our sample were evaluated at different points in their career. This allows us to identify the effect of evaluation on performance separate from any gains that come from increased experience. We return to this topic in our discussion of empirical strategy. Third, the delay in first evaluation allows us to observe the achievement gains of these teachers' students in classes the teachers taught before TES evaluation. As we describe in the next section, these before-evaluation years serve as our counterfactual in a teacher fixed effects estimation strategy.

ing" evaluation; this translates into a situation where teachers are effectively not evaluated in non-TES years. As described in this section the full TES program is quite different. In this paper we focus on the full TES evaluation, and all references to "evaluation" are to that system.

[8] Some teachers in our sample volunteered to be evaluated years before their scheduled participation. We return to the effect of these off-schedule teachers on our estimate in the results section. But, in short, their inclusion does not dramatically affect our estimates.

Additionally, this paper focuses on math teachers in grades four through eight. For most other subjects and grades, student achievement measures are simply not available. Students are tested in reading but empirical research frequently finds less teacher-driven variation in reaching achievement compared to math (Hanushek and Rivkin 2010), and ultimately this is the case for the present analysis as well. While not the focus of this paper, we discuss reading results in a later section and present reading results in online Appendix Table A1.

Data provided by the CPS identify the year(s) in which a teacher was evaluated by TES, the dates when each observation occurred, and the scores. We combine these TES data with additional administrative data provided by the district that allow us to match teachers to students and student test scores.

Panel A of Table 2 contrasts descriptive characteristics of the teachers in our analysis sample (row 1) with the remaining fourth through eighth grade math teachers and students in Cincinnati during this period but not included in our sample (row 2). The third row provides a test of the difference in means or proportions. As expected given its construction, our sample is more experienced. Indeed, the one year mean difference understates the contrast: 66.5 percent of the analysis sample is teachers with 10 to 19 years experience compared to 29.3 percent of the rest of district. Analysis sample teachers are also more likely to have a graduate degree and be National Board certified, two characteristics correlated with experience.

The leftmost columns of Table 3 similarly compare characteristics for the students in our analysis sample (column 1) with their peers in the district taught by other teachers (column 2). Column 3 provides a test of the difference in means or proportions. The analysis sample is weighted toward fourth through sixth grade classes, and analysis students may be slightly higher-achieving than the district average.[9]

While the observable student and teacher differences are not dramatic, nearly all are statistically significant. These differences reinforce the limits on generalizing to more- or less-experienced teachers, but are not necessarily surprising. Researchers have documented large differences in the students assigned to more experienced teachers (Clotfelter, Ladd, and Vigdor 2005, 2006). The remainder of the information in Tables 2 and 3 explore whether these observable teacher characteristics are related to treatment status. We return to this discussion after presenting our empirical strategy in the next section.

III. Empirical Strategy

Our objective is to estimate the extent to which subjective performance evaluation of CPS teachers impacts teacher productivity. We employ a teacher fixed effects approach to estimate the model of student math achievement described by equation 1:

$$(1)\quad A_{ijgt} = f(\textit{TES evaluation}_{jt}) + g(\textit{Experience}_{jt}) + \mathbf{A}_{ijg(t-1)}\boldsymbol{\alpha} + \mathbf{X}_{ijgt}\boldsymbol{\beta} + \mu_j + \theta_{gt} + \varepsilon_{ijgt},$$

[9] The district mean for test scores in Table 3 is not zero because we include only students who have baseline and outcome math scores. Students missing test scores are mostly those who move in and out of the district, and that enrollment instability is associated with lower test scores.

Table 2—Teacher Characteristics

	Years experience	Graduate degree	National Board certified	Female	African-American	White
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Teacher characteristics for analysis sample and remainder of district						
Analysis sample ($n = 105$)	13.75	0.7195	0.0670	0.8536	0.3903	0.5579
	(3.94)					
Remainder of district ($n = 560$)	12.99	0.5443	0.0324	0.8293	0.2676	0.7101
	(6.60)					
Difference *t*-test *p*-value	0.000	0.000	0.004	0.295	0.000	0.000
Panel B. Teacher characteristics by actual evaluation year[a]						
2009–2010 actual evaluation	14.444	0.778	0.222	0.778	0.333	0.667
Difference relative to 2009–2010						
2003–2004	−9.444***	−0.278	0.028	0.222	0.167	−0.167
	(2.576)	(0.277)	(0.164)	(0.208)	(0.294)	(0.303)
2004–2005	−8.444***	0.022	−0.222	0.222	−0.133	−0.067
	(2.391)	(0.257)	(0.152)	(0.193)	(0.273)	(0.282)
2005–2006	−4.444**	0.000	−0.167	0.111	0.000	0.000
	(1.750)	(0.188)	(0.111)	(0.142)	(0.200)	(0.206)
2006–2007	−1.778	−0.044	−0.089	0.222	0.000	−0.067
	(1.807)	(0.194)	(0.115)	(0.146)	(0.207)	(0.213)
2007–2008	−2.044	−0.138	−0.222**	−0.018	0.147	−0.147
	(1.666)	(0.179)	(0.106)	(0.135)	(0.191)	(0.196)
2008–2009	3.556*	0.022	−0.122	0.022	−0.133	0.033
	(1.970)	(0.212)	(0.125)	(0.159)	(0.225)	(0.232)
Joint *F*-test *p*-value	0.000	0.860	0.303	0.339	0.736	0.948
Panel C. Teacher characteristics by scheduled evaluation year[a]						
2009–2010 scheduled evaluation	15.714	0.857	0.286	0.857	0.000	1.000
Difference relative to 2009–2010						
2005–2006	−5.714***	−0.048	−0.238**	0.048	0.476**	−0.524**
	(1.800)	(0.190)	(0.107)	(0.150)	(0.208)	(0.209)
2006–2007	−5.893***	−0.214	−0.179	0.071	0.429**	−0.500**
	(1.743)	(0.184)	(0.104)	(0.145)	(0.202)	(0.202)
2007–2008	−5.048***	−0.090	−0.252**	−0.024	0.500**	−0.533***
	(1.731)	(0.183)	(0.103)	(0.144)	(0.200)	(0.201)
2008–2009	1.654	−0.068	−0.286***	−0.068	0.211	−0.211
	(1.823)	(0.193)	(0.109)	(0.152)	(0.211)	(0.211)
Joint *F*-test *p*-value	0.000	0.612	0.086	0.663	0.050	0.018

Notes: Teachers of fourth through eighth grade math students, 2003–2004 through 2009–2010. Standard errors in parentheses.

[a] Each column reports a separate teacher-level regression of the characteristic on indicators for evaluation year.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

TABLE 3—STUDENT CHARACTERISTICS

	Analysis sample	Remainder of district	Difference *t*-test *p*-value	Regression-adjusted[a] difference relative to all years prior to evaluation: Year of evaluation ($t = \tau$)	Regression-adjusted[a] difference relative to all years prior to evaluation: All years after ($t > \tau$)
	(1)	(2)	(3)	(4)	(5)
Baseline math test score	0.0846	0.0713	0.193	−0.012	−0.037
	(0.954)	(1.009)		(0.044)	(0.061)
Baseline reading test score	0.1104	0.0614	0.000	0.013	−0.011
	(0.949)	(0.999)		(0.052)	(0.057)
Female	52.15	49.51	0.000	0.000	0.006
				(0.014)	(0.017)
African-American	71.58	69.79	0.000	−0.047	0.007
				(0.037)	(0.029)
White	21.33	23.87	0.000	0.047	−0.001
				(0.036)	(0.024)
Special education	16.84	19.29	0.000	0.017	0.026
				(0.016)	(0.018)
English language learner	3.41	2.49	0.000	0.000	0.011
				(0.006)	(0.008)
Gifted and talented	12.65	9.77	0.000	0.005	−0.010
				(0.022)	(0.021)
Retained in grade	0.65	1.01	0.000	−0.002	−0.001
				(0.004)	(0.003)
Grade level					
4th	27.97	22.11	0.000		
5th	26.01	16.64	0.000		
6th	16.84	13.44	0.000		
7th	17.31	25.45	0.000		
8th	11.88	22.37	0.000		
Student-year observations	14,331	44,952			

Notes: Fourth through eighth grade math students, 2003–2004 through 2009–2010. Includes only students with both baseline and outcome math test scores.

[a] Each row reports a separate teacher fixed effects regression of the student characteristic on grade-by-year fixed effects, a quadratic in teacher experience, and the indicated covariates. Clustered (teacher) standard errors in parentheses. In each case, the sample is composed of 105 teachers and 14,331 student-by-year observations, except baseline reading score, where there are 12,207 student observations.

where A_{ijgt} represents the end-of-year math test score[10,11] of student i taught by teacher j in grade g and school year t. In all estimates, standard errors are clustered by teacher.

[10] All test scores have been standardized (mean zero, standard deviation one) by grade and year. Between 2002–2003 and 2009–2010 Cincinnati students, in general, took end-of-year exams in reading and math in third through eighth grades. Our analysis sample will exclude some entire grade-by-year cohorts for whom the state of Ohio did not administer a test in school year t or $t - 1$.

[11] An alternative standardization approach would use only the distribution of our analysis sample instead of the entire district. The average experience level of our analysis sample, which is comprised of a fixed set of hire-year cohorts, will be steadily rising in the district distribution of teacher experience. If the changes over time in relative experience are dramatic then the district standardization could produce a positive time trend since new hires are generally of lower ability in their first few years on the job. In analyses not presented we repeat all our main results using this alternative standardization and do not find substantial differences.

The key empirical challenge is to separately identify (i) the effect of subjective performance assessment via TES evaluation on productivity in years during and after evaluation, $f(TES\ evaluation_{jt})$; (ii) the effect of increasing job experience, $g(Experience_{jt})$; and (iii) the secular trends in student test scores, θ_{gt}, from year to year and grade to grade. For any individual teacher across years, these three determinants of student test scores—year relative to TES evaluation year, years of experience, and school year—will be collinear. Identification requires some parameter restriction(s) for two of the three determinants. Given our use of standardized test scores as a measure of achievement, we maintain the inclusion of grade-by-year fixed effects in all estimates to account for θ_{gt}, and thus must make some restriction in both f and g.

Most of the estimates we present use a simple parameterization for time relative to TES participation, specifically

$$f(TES\ evaluation_{jt}) = \delta_1 \mathbf{1}\{t = \tau_j\}_{jt} + \delta_2 \mathbf{1}\{t > \tau_j\}_{jt}, \tag{2}$$

where τ_j is the school year during which teacher j was evaluated. School years before TES evaluation are the omitted category. Thus, δ_1 captures the gain (loss) in achievement of students taught by a teacher during his TES evaluation year compared to students he taught before being evaluated, and δ_2 captures the gain (loss) of students taught in years after TES evaluation compared to before.

Almost all the results presented in this paper control for a quadratic in teacher experience, but with teacher experience capped at 20 years.

$$g(Experience_{jt}) = \gamma_1 \widetilde{Experience}_{jt} + \gamma_2 \widetilde{Experience}^2_{jt} \tag{3}$$

$$where\ \widetilde{Experience}_{jt} = \begin{cases} Experience_{jt}, & if\ Experience_{jt} < 20 \\ 20, & if\ Experience_{jt} \geq 20 \end{cases}$$

This approach follows Rockoff (2004), though our cap is higher given our more experienced sample. In results not presented here, we repeat our main analyses in two ways, one using a cap at 15 years, and a second that replaces the linear $\widetilde{Experience}_{jt}$ with indicator variables for each discrete value of $\widetilde{Experience}_{jt}$; our results are essentially unchanged in these alternative specifications.

Much empirical evidence finds the returns to experience are greatest in the first five or ten years of a teacher's career (Rockoff 2004; Staiger and Rockoff 2010). This evidence suggests omitting experience may not introduce much bias in our estimates given the experience profile of the teachers in our analysis sample. And, as reported later, our preferred estimates do not change when the experience controls are excluded. Nevertheless, we maintain the experience controls partly because of more recent evidence that finds experience gains beyond ten years (Papay and Kraft 2010; Wiswall 2011).

The use of the within-teacher over-time variation is preferable here for at least two reasons. First, existing evidence suggests that both inexperienced and experienced teachers vary greatly in their ability to promote student achievement (Hanushek and Rivkin 2010). To the extent that teacher ability is correlated with participation

in and the timing of TES evaluation (e.g., through differential attrition from the district, or volunteering for the program) simple cross-sectional estimates would be biased. Second, the teacher fixed effects will account for time-invariant, nonrandom differences in the assignment of students to specific teachers. Some teachers may be asked to teach classes year after year with high (low) potential for achievement gains (e.g., through principal favoritism, or school assignment).

Not all the dynamics of student-teacher assignment need be time-invariant, however. To account for variation in students assigned to a given teacher from year to year, in equation (1) we control for observable student characteristics. Most notably, $\mathbf{A}_{ijg(t-1)}$ includes student i's prior achievement, as measured by the year $t-1$ test, the effect of which is allowed to vary by grade.[12] The vector $\mathbf{X}_{ijgt}$ includes separate indicators for student gender, racial/ethnic subgroup, special education classification, gifted classification, English proficiency classification, and whether the student was retained in grade.

Our teacher fixed effects estimation approach will provide unbiased estimates of δ_1 and δ_2 if, for a given teacher, the timing of her evaluation is unrelated to student achievement trends not captured by her assigned student's test scores and other observable characteristics, the average returns to experience, or district-wide secular trends. This key identifying assumption would be violated, for example, if teachers were systematically assigned unobservably better (worse) students during their evaluation year or in the years following. It would also be violated if evaluation coincided with an individual performance trend unrelated to evaluation per se.

After presenting our main estimates, we provide evidence that the results are not driven by a preexisting trend in teacher performance with a nonparametric event study and robustness checks through alternative specifications of equation (2). We also address the potential for attrition bias.

Before moving on to results, however, we present evidence that, at least based on observables, student assignment to teachers was unrelated to a teacher's evaluation status, and that teacher observables are also largely uncorrelated with the timing of evaluation. The rightmost columns of Table 3 report coefficients from a series of regressions predicting each student covariate ($A_{ijg(t-1)}$ and the elements of $\mathbf{X}_{ijgt}$) as a function of teacher TES status. The specification is identical to that of our preferred specification described in equation (1) except for omitting the student covariates themselves. None of the estimates for δ_1 and δ_2 are statistically significant, and most point estimates were near zero. Thus, despite the variation across teachers in assigned students, we observe little variation within teachers over time.[13]

The bottom two panels of Table 2 examine potential differences in the characteristics of analysis sample teachers by year of evaluation, where the characteristics

[12] When the baseline score was missing for a student in 13.1 percent of observations, we imputed with the grade-by-year mean, and included an indicator for missing baseline score. Our estimates are robust to excluding students with missing baseline test scores.

[13] In a second approach to this question, following Rothstein (2010) and others, we test whether a student's future teacher, in year $t+1$, is related to current achievement growth, in year t. In our present case we are interested in the future teacher's TES treatment status. We repeat the specification in Table 4 column 1 except that we replace the variables of interest "year of participation" and "all years after" with their equivalents for the teacher who taught student i in year $t+1$. Again we find no evidence of differential assignment. We estimate that student achievement growth is negative, though not statistically significant, in the year just prior to being assigned a teacher under evaluation or postevaluation.

Table 4—Math Achievement Gains for Students Taught during and after Teacher Evaluation

				Quartile interaction		
				Overall TES score	TES gain during evaluation	Pre-TES test-score value-added
	(1)	(2)	(3)	(4)	(5)	(6)
School year relative to year of TES evaluation (*all other years prior omitted*)						
Year of evaluation ($t = \tau$)	0.052	0.057	0.061	0.047	0.060	0.057
	(0.036)	(0.036)	(0.045)	(0.036)	(0.036)	(0.038)
All years after ($t > \tau$)	0.112**	0.117**	0.075			
	(0.048)	(0.048)	(0.057)			
All years after × bottom quartile				0.180*	0.054	0.314***
				(0.098)	(0.078)	(0.083)
All years after × 2nd quartile				0.149**	0.042	0.092*
				(0.066)	(0.094)	(0.048)
All years after × 3rd quartile				0.040	0.174***	0.043
				(0.098)	(0.065)	(0.115)
All years after × top quartile				0.088	0.199***	0.011
				(0.059)	(0.072)	(0.077)
Teacher experience quadratic	Yes			Yes	Yes	Yes
Student covariates	Yes	Yes		Yes	Yes	Yes

Notes: Each column reports a separate teacher fixed effects estimation of student standardized (by grade and year) math test score as a function of grade-by-year fixed effects, and the indicated covariates. Student covariates include prior year achievement (main effect, interaction with grade level, and indicator for missing value) and indicators for gender, race/ethnicity subgroup, special education classification, English language learner classification, gifted and talented classification, and students retained in grade. Clustered (teacher) standard errors in parentheses. In each case the sample is composed of 105 teachers and 14,331 student-by-year observations.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

are given by the global column headers. The mean of each characteristic for 2009–2010 is given in the first row of each panel, and all other rows are differences from that mean as estimated by a simple least-squares regression. The last row reports the results of a test of joint significance. Panel B is calculated using the actual year of evaluation, and, with the expected exception of experience, there are almost no significant differences. Panel C uses the scheduled year of evaluation, instead of the actual year, and the pattern is similar except that we find evidence of differences in race.

IV. Results

Table 4, column 1 reports the coefficients of interest from our preferred specification: equation (1) with the simple parameterization of the TES evaluation treatment measures in equation (2). We estimate that the average math teacher's students score 0.112 standard deviations higher in years after the teacher has undergone an evaluation compared to how her students scored in years before evaluation. Students also appear to score higher during the year of the subjective evaluation (0.052 standard deviations), but in this specification that difference is not statistically significant.

These results are robust to omitting the teacher experience controls (equation (3)), as shown in column 2. In results not presented here, we find this robustness is also true for other forms of experience controls: i.e., capping at 15 instead of 20 years, and a series of dummy variables for each year of experience. When the student covariates, including prior achievement, are also omitted, as shown in column 3, the estimates are somewhat less precise and the coefficient for "all years after" is somewhat smaller.

The right side of Table 4 presents evidence that the effects of going through subjective evaluation in the TES system are not uniform. The improvement in teacher performance from before to after evaluation is larger for teachers who received relatively low TES scores (column 4), teachers whose TES scores grew the most during the TES year (column 5), and teachers with relatively low test-score value-added prior to TES (column 6).[14] For each of these three measures of performance we separated teachers into quartiles, and interacted the quartiles with the indicator for years after TES.[15] Since we do not have strictly pretreatment measures of TES scores, we focus here on heterogeneity in the years after effects.

This pattern of heterogeneity is consistent with a causal interpretation of our estimates. While all teachers were evaluated, not all teachers had the same potential for growth ex ante. Assuming subjective evaluation causes growth in productivity, that growth should vary from individual to individual as a function of ex ante potential for growth. The results in Table 4, columns 4–6 demonstrate just this kind of pattern. Teachers who prior to evaluation generated relatively little value-added to student test scores saw the largest productivity gains in the years following evaluation (column 6).[16] Gains were also larger among teachers whose TES evaluation scores were relatively low (column 4), and among teachers whose TES scores increased the most across the four classroom observations that occurred during the evaluation year (column 5).

Additionally, there is some evidence that the effects are larger for teachers who specialize in teaching math, as opposed to teaching math and reading in self-contained classrooms. To analyze this question, we fit separate models for the 50 teachers in our sample who teach only math and the 55 teachers who teach math and reading in self-contained classrooms. For the former, student achievement in the years after evaluation is 0.11 standard deviations higher. For the latter, the gain is 0.04 standard deviations but is not statistically significant.

In contrast to the results for math achievement, which we focus on in this paper, we do not find statistically significant differences for reading teachers and reading achievement. With a specification matching Table 4 column 1, the estimates for

[14] Throughout the paper "test-score value-added" is the estimated total effect, including the returns to experience and differences in skill and practice, of an individual teacher on student achievement (see Hanushek and Rivkin 2010 for an introduction and review of empirical estimates). The metric is measured in student test score standard deviations. To form the test-score value-added estimates for this paper we estimate a specification like equation (1); however, we omit the experience and TES controls, add school fixed effects and class random effects, and use all teachers and students in the district. The value-added estimates are predicted teacher random effects, μ_j, shrunken to account for measurement error.

[15] The TES scores are based on two dozen teaching practices and are collectively known as TES Domains 2 and 3. See Kane et al. (2011) for more information about the process, rubric, and scores. TES score growth is the change in overall TES score from the first to the last classroom observation during the TES year.

[16] While we cannot rule out some role for mean reversion as an explanation for the results in Table 4 column 6, the event study depicted in Figure 1 makes this explanation unlikely. We do not see similar reversion at the other end of the distribution, however.

TABLE 5—ALTERNATIVE SPECIFICATIONS OF EVALUATION TREATMENT

	Variations in functional form			Scheduled evaluation	2SLS
	(1)	(2)	(3)	(4)	(5)
School year relative to year of TES evaluation (*all other years prior omitted*)					
Year immediately prior ($t = \tau - 1$)	0.045				
	(0.043)				
Year of evaluation ($t = \tau$)	0.087*	0.064*	0.064*	0.027	0.028
	(0.048)	(0.037)	(0.038)	(0.022)	(0.059)
All years after ($t > \tau$)	0.157**			0.069**	0.128**
	(0.062)			(0.031)	(0.060)
First year after ($t = \tau + 1$)		0.112**	0.113**		
		(0.048)	(0.056)		
Two or more years after ($t \geq \tau + 2$)		0.157**			
		(0.067)			
Second year after ($t = \tau + 2$)			0.158**		
			(0.072)		
Three or more years after ($t \geq \tau + 3$)			0.161		
			(0.106)		
First-stage *F*-statistic on excluded instruments					
Year of evaluation					864
All years after					1,439

Notes: Each column reports a separate teacher fixed effects estimation of student standardized math test score. Additional covariates as described in table 4 note including teacher experience and student covariates. Two-stage least-squares (2SLS) estimates use scheduled participation, based on the TES phase-in schedule, as instruments for actual participation. Clustered (teacher) standard errors in parentheses. In each case the sample is composed of 105 teachers and 14,331 student-by-year observations.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

reading are −0.009 (0.036) in the year of evaluation, and −0.050 (0.038) in the years following evaluation. In the next section we briefly discuss why the lack of results in reading may not be surprising. Results for both reading and specialized versus self-contained math teachers are available in the online Appendix Table A1.

A. *Alternative Specifications*

In Table 5 and Figure 1 we vary the specification of $f(\textit{TES evaluation}_{jt})$ to explore the robustness of our main estimates. Most notably, the estimates presented in Table 4 may be biased by some preexisting upward trend in teacher performance unrelated to experience growth. We address the trend question first, and then a second related question of specification.

In Figure 1 we plot the average student test score gain, and confidence interval, in each year relative to the TES evaluation year. Each point is the coefficient from a regression, similar to equation (1), where $f(\textit{TES evaluation}_{jt})$ is a vector of indicator variables for each year relative to evaluation $(t - \tau)$. This specification excludes teacher experience controls for the identification reasons discussed earlier. The vertical axis is normalized to zero in the year immediately prior to evaluation, mathematically $t - \tau = -1$. While the estimates are noisy, there is no clear trend in the

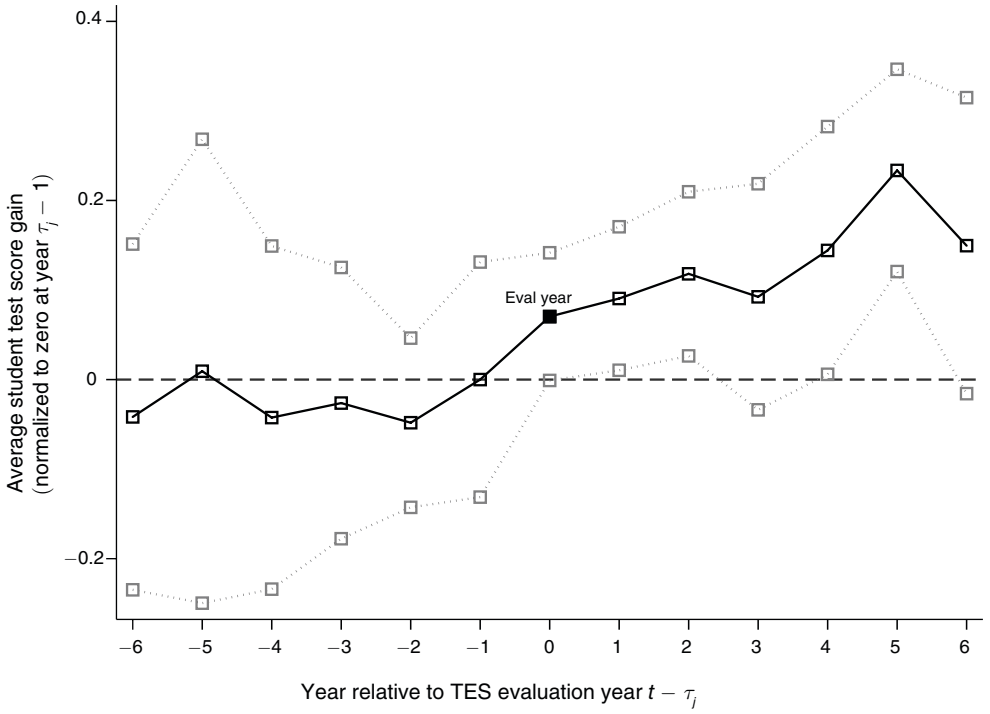

FIGURE 1. TEACHER PERFORMANCE RELATIVE TO YEAR OF EVALUATION

Notes: Dotted lines represent 95 percent confidence interval. Sample composed of 14,331 students and 105 teachers.

years prior to evaluation—a pattern that supports our parameterization that groups "all years prior" as a single omitted condition in equation (2). Figure 1 also shows the gains after evaluation, and to some extent during evaluation, that are reflected in the estimates in Table 4.[17]

In the online Appendix we provide similar figures for subsamples of teachers, including teachers in the bottom quartile of preevaluation test-score value-added and teachers in the bottom two quartiles of that distribution. As in the full sample the figures are consistent with Table 4 and do not demonstrate a preexisting trend, though the estimates are noisier given the smaller samples.

While Figure 1 provides important evidence against a preexisting trend, it does not account for experience. Thus, in Table 5 columns 1–3 we report estimates from marginal variations on our preferred specification while controlling for experience. In column 1 we include the year immediately prior to evaluation separately; i.e., $\mathbf{1}\{t = \tau_j - 1\}_{jt}$; this is one test of whether teachers were on an upward (downward) trend heading into their evaluation year. The results suggest that teachers may have been on an upward trajectory not captured by our measure of experience, but we cannot rule out that the slight trend we estimate is the result of chance. In column 2

[17] In Table A2, available in the online Appendix, we report estimates of equation (1) with a specification of $f(TES\ evaluation_{jt})$ mirroring Figure 1, but without teacher experience controls. The estimated coefficients for each year relative to evaluation follow the pattern represented in Figure 1.

we include an indicator for the year immediately following TES participation, $\mathbf{1}\{t = \tau_j + 1\}_{jt}$, separately from two or more years after, $\mathbf{1}\{t \geq \tau_j + 2\}_{jt}$. In column 3 we further separate out the second year following participation. In each case the pattern of results is not much different from the simple specification in Table 4 column 1. The estimates become less precise, however, as the sample in the "or more years" category shrinks.

We turn now to a second question. In practice, not all teachers were evaluated during their assigned phase-in schedule year: 63.8 percent were evaluated on-schedule.[18] The program's administrators describe most deviations from the phase-in schedule as administrative and unrelated to a teacher or student performance; one teacher might participate a year late because he was on leave the scheduled year, while another might participate a year early to fill out an evaluator's case load. Nevertheless, our estimates in Table 4 that rely on actual evaluation year could be biased by teachers strategically timing their evaluation (e.g., through volunteering) to coincide with a positive—but unobservable in the data—shock in the type of students they are assigned.[19] Such a source of bias is less plausible for our estimates of gains in the years following evaluation: teachers would have to time their evaluation relative to a positive shock that persists into the future.

In column 4 of Table 5 we report estimates identical to Table 4, column 1 except that we use scheduled evaluation year instead of actual evaluation year. In these alternative estimates, teacher performance is still meaningfully, and statistically significantly, better in years following scheduled, as opposed to actual, evaluation. The smaller magnitude is not necessarily unexpected. When there is a difference, actual evaluation almost always precedes scheduled evaluation; thus, if evaluation does have a lasting positive impact, the counterfactual measure (i.e., performance in years prior to *scheduled* evaluation) would be biased upward, resulting in effect estimates that are too small.

Column 5 of Table 5 reports two-stage least squares (2SLS) estimates instrumenting for actual participation timing using the scheduled timing. If our main effects were driven by positive bias from strategic timing of evaluation then the effect among those who comply with the schedule (i.e., the 2SLS estimate) should be smaller; this is not the case for the effect in years following evaluation.

B. *Attrition*

Since TES is a performance *evaluation* program, teachers who scored low may be more likely to stop teaching in the district after their TES year and thus attrit from our sample. Teachers who scored *high* may also be more likely to attrit since high TES scores are explicitly used by the district in some promotion decisions. But attrition correlated with high or low *levels* of performance is not necessarily a problem given our teacher fixed effects strategy; the estimates in this paper measure the

[18] An additional 10.5 percent volunteered to be evaluated years ahead of schedule. Teachers did so primarily to post evaluation scores so that they could be considered for promotions within districts that require high TES scores. The results presented in Table 4 are robust to excluding these early volunteers.

[19] As shown in Table 3 we find no significant differences in observable student characteristics correlated with actual evaluation timing.

average teacher improvement (decline) in years after evaluation compared to each teacher's own performance in years before evaluation. Thus, our estimates would be biased only if attrition is correlated with the trajectory of change in teacher effectiveness after evaluation.[20]

After their TES evaluation year, 96 percent of teachers in our analysis sample were still teaching in CPS; however, 9.5 percent were teaching in some grade or subject combination outside our fourth–eighth grade math data. These rates are consistent with the district's patterns of turnover generally, regardless of evaluation requirements. [21]

Still, whatever the causes for the attrition we observe, that attrition may nevertheless be correlated with treatment and the outcome of interest—the improvement or decline in performance as a result of TES evaluation. First, we measure the extent to which attrition is correlated with the timing of TES evaluation. In the top panel of Table 6 we report estimates from teacher-by-year regressions predicting a binary outcome equal to 1 if the teacher j is observed in the sample in year t. Columns 1 and 2 are based on teachers' actual evaluation timing; columns 3 and 4 are based on scheduled timing.[22] The pattern of negative estimates on the "year of evaluation" and the "all years after" coefficients in Table 6 offer some evidence that teachers may be less likely to be observed in the data during and after evaluation relative to the before-evaluation years. None of the differences in the probability of being in the sample are statistically significant, however. On the other hand, the coefficients on lagged test-score value-added indicate that more effective teachers are less likely to attrit in any year, but again, attrition based on level of performance per se does not threaten our identification.

In the bottom panel of Table 6 we compare our preferred estimates (repeated in column 1) to two alternatives. In column 3 we restrict the sample to teachers whom we observe both before and after evaluation, thus excluding any attriters ex ante.[23] Under this restriction, the pattern of gains remains consistent; if anything, the estimates suggest larger gains. The online Appendix provides a figure similar to Figure 1 for this subsample; the pattern of results is consistent. This exercise does not eliminate attrition bias for our main estimates, but suggests our identification is not driven by attriters.

Columns 2 and 4 present weighted versions of columns 1 and 3, respectively. Observations are weighted by the inverse predicted probability of remaining in the

[20] Evaluation may also generate anticipatory attrition. That is, teachers' decisions to leave the district or switch to other subjects and grades may have been influenced by the prospect of subjective evaluation in the TES system. If such anticipatory attrition occurred, our within-teachers strategy will still produce internally valid estimates of the effect on the "treated" teachers, but the attrition would suggest potential general equilibrium effects of subjective evaluation as a whole. Of the 140 teachers hired between 1993–1994 and 1999–2000 who taught math in grades four through eight, 11.4 percent stopped teaching in the district before their TES evaluation year.

[21] In any given year, 90 percent of math 4–8 teachers will return to teach somewhere in the district the following year; 60 percent will return to teach math 4–8. Among teachers hired 1993–1994 to 1999–2000, 60 percent teach math 4–8 in consecutive years. For teachers hired 1986–1987 to 1992–1993—teachers not subject to TES evaluation—the rate is 61 percent. For teachers hired 2000–2001 to 2003–2004—subject to evaluation in their first year—the rate is 58 percent.

[22] The sample for this analysis includes teachers who stopped working in the district before their scheduled TES evaluation. For those teachers in columns 1 and 2 we use the scheduled evaluation year.

[23] While this subsample is composed of teachers who are observed in multiple years throughout the study period both before and after evaluation, not all teachers are observed in all years. During the study period the state of Ohio did not test math in all grades all years, and to be "observed" at year t a teacher must have a class with test scores in years t and $t-1$. Thus, while the traditional approach to such an analysis is to use a truly balanced panel, we use this pseudo-balanced panel.

TABLE 6—ATTRITION RELATED ESTIMATES

	Actual evaluation		Scheduled evaluation	
	(1)	(2)	(3)	(4)
Panel A. Modeling teacher attrition				
Dependent variable: Teacher j observed in sample in year t				
School year relative to year of TES evaluation (all other years prior omitted)				
Year of evaluation ($t = \tau$)	0.010	−0.007	−0.059	−0.068
	(0.051)	(0.050)	(0.054)	(0.051)
All years after ($t > \tau$)	−0.064	−0.069	−0.123	−0.098
	(0.051)	(0.042)	(0.078)	(0.062)
Test-score value-added at time $t - 1$		0.105***		0.104***
× has value-added measure at time $t - 1$		(0.024)		(0.023)
Experience quadratic, graduate degree, and salary at time $t - 1$ controls		Yes		Yes
Teacher clusters	140	140	140	140
Teach-year observations	980	980	980	980

	Full sample		Observed before and after evaluation	
	(1)	(2)	(3)	(4)
Panel B. Alternative effect estimates				
School year relative to year of TES evaluation (all other years prior omitted)				
Year of evaluation ($t = \tau$)	0.052	0.044	0.103**	0.101**
	(0.036)	(0.036)	(0.043)	(0.044)
All years after ($t > \tau$)	0.112**	0.105**	0.153***	0.150***
	(0.048)	(0.048)	(0.051)	(0.051)
Weighted: inverse predicted probability of remaining in sample after evaluation		Yes		Yes
Teacher clusters	105	105	82	82
Student-year observations	14,331	14,331	12,453	12,453

Notes: Panel A are teacher fixed effects linear probability models predicting presence in the sample as a function of the covariates listed, year fixed effects, and an indicator for the availability of a lagged value-added score. Panel B teacher fixed effects regressions are as described in the table 4 note, including teacher experience and student level covariates. See text for description of weights estimation. Clustered (teacher) standard errors in parentheses.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

data (not-attriting).[24] The intuition behind this approach is that a teacher with, for example, a 0.25 probability of persisting who actually does persist will be weighted up to represent four teachers on the expectation that three others with a probability of 0.25 are not observed. Of course, some unobserved factor(s) may have caused one teacher with a probability of 0.25 to persist while others with the same probability did not. In the end, the weighted and unweighted estimates in in the bottom panel of Table 6 are very similar.

[24] The predicted probabilities underlying the weights come from an auxiliary probit regression similar to the top panel of Table 6. Using teacher-level data we predict presence in the sample after TES evaluation as a function of TES score, value-added prior to evaluation, a quadratic in teacher experience, salary in 2004, gender, degree-level, and national board certification. We thank Ken Chay for suggesting this analysis. Any errors are our own.

Last, we provide a back-of-the-envelope calculation for how our main estimates would be changed by different assumptions about the effect of TES evaluation on the attriters. Assume that the performance of attriters was unaffected by evaluation: their unobserved postevaluation performance was (or would have been) the same as their preevaluation performance. Thus, if we could estimate equation (1) using just the sample of attriters, we would expect the coefficient on "all years after" to be zero. Under this first assumption, our updated estimate of the postevaluation performance gain would be about 0.087.[25] Alternatively, the performance of the attriters may have been negatively affected by TES participation. But the point estimate in Table 4, column 1 would be zero only if the negative effect on attriters was roughly 3.5 times as large, in absolute value, as the current positive estimate.

V. Discussion

The estimates presented here—greater teacher productivity as measured by student achievement gains in years following TES evaluation—strongly suggest that teachers develop skill or otherwise change their behavior in a *lasting* manner as a result of undergoing subjective performance evaluation in the TES system. Imagine two students taught by the same teacher in different years who both begin the year at the fiftieth percentile of math achievement. The student taught after the teacher went through comprehensive TES evaluation would score about 4.5 percentile points higher at the end of the year than the student taught before the teacher went through the evaluation.

Such changes are consistent with a model, discussed at the outset, where teachers learn new information about their own performance during the evaluation and subsequently develop new skills. New information, the key mechanism, is potentially created by the formal scoring and feedback routines of TES, as well as the teacher self-reflection required in TES evaluation and the increased opportunities for conversations around effective teaching practice in a TES evaluation environment. Moreover, two features of this study—the analysis sample of experienced teachers and Cincinnati's use of peer evaluators—may lend greater saliency to these hypothesized mechanisms. First, the teachers we study experienced their first rigorous evaluation after 8 to 17 years on the job. Thus, they may have been particularly receptive to and in need of information on their performance. If, by contrast, teachers were evaluated every school year (as they are in a new but similar program in Washington, DC), the effect resulting from each subsequent year's evaluation might well be smaller. Second, Cincinnati's use of peer evaluators may result in teachers being more receptive to feedback from their subjective evaluation relative to how they might view this information were it coming solely from their supervising principals.[26]

[25] This is simply the estimate in Table 4 column 1, 0.112, multiplied by 0.78, the proportion observed. If the standard error remained roughly the same, then a difference of 0.087 would be significant at the 10 percent level. We thank Caroline Hoxby for suggesting this analysis, any errors are our own.

[26] The performance improvements in years after subjective evaluation could, under one alternative hypothesis, continue to be a response to direct evaluation. Imagine, for example, that school administrators give greater scrutiny to the work of teachers who recently scored low on their TES evaluation. Teachers may respond to that scrutiny by boosting their efforts even though they are not subject to formal evaluation in the postevaluation years.

We cannot, however, say what teachers changed about their behavior or practice, nor which changes were most important to student achievement growth. While there is evidence that following the TES rubric per se should help (Kane et al. 2011), following the rubric is only one possible mechanism. Alternatively, the general peer- and self-scrutiny may have uncovered opportunities for improvement in areas not addressed by the TES rubric.

Teachers also appear to generate higher test score gains during the year they are being evaluated, though these estimates, while consistently positive, are not always significant. These improvements during the evaluation could represent the beginning of the changes seen in years following evaluation, or they could be the result of simple incentives to try harder during evaluation, or some combination of both.

While our focus in this paper has been on math achievement, however, in similar analyses of reading achievement we do not find significant differences in student achievement growth associated with TES evaluation. Many studies find less variation in teachers' effects on reading achievement compared to the variation in teachers' effects on math achievement (Hanushek and Rivkin 2010). Some have hypothesized that these smaller reading teacher differences could arise because students learn reading in many in- and out-of-school settings (e.g., at home) that are outside of a formal reading class. If teachers have less influence on reading achievement variation, then changes in teacher practices would have smaller returns.

Before concluding we briefly discuss the magnitude of the gains estimated here, and the costs associated with evaluation in Cincinnati. A natural comparison for calibrating the size of these effects would be teacher professional development programs (in-service training, often delivered in formal classroom settings). Unfortunately, despite the substantial budgets allocated to such programs, there is little rigorous evidence on their effects (see Yoon et al. 2007 for an extensive review).

There are, however, other estimates from the general literature on teacher human capital development that can be used for comparison. First, among extant evidence, the largest gains in teacher effectiveness appear to occur as teachers gain on-the-job experience in the first three to five years. Rockoff (2004) reports gains of about 0.10 student standard deviations over the first two years of teaching when effectiveness is measured by math computation gains; when measured by math concepts, the gains are about half as big and not statistically significant. Second, Jackson and Bruegmann (2009) study the effect of working around more effective colleagues, and find that better teacher peers improves a teacher's own performance. A one standard deviation increase in teacher-peer quality was associated with a 0.04 student standard deviation increase in math achievement.

The TES evaluation program carries two important types of cost: (i) the salaries of TES evaluators and staff, and other direct program costs; and (ii) the opportunity cost in student achievement terms incurred by allocating effective, experienced classroom teachers to evaluator roles. First, it should not be surprising that the budget expenditure is relatively large given the atypically intense TES approach. From 2004–2005 to 2009–2010 the district budget directly allocated between $1.8 million and $2.1 million per year to the TES program, or about $7,500 per teacher evaluated (CPS 2010). Over 90 percent of this cost is associated with evaluator salaries.

Second, the classroom teaching positions vacated by individuals selected to be peer evaluators will, presumably, be filled with less-effective, likely novice teachers.[27] Viewing evaluation from a human capital development perspective, the net loss in productivity—the production of student achievement—from these substitutions is a central cost of the investment (Becker 1962).

Assume that TES evaluators are drawn from the top quartile of the district's teacher effectiveness distribution, and their replacements from the bottom quartile. In Cincinnati, a student in the classroom of a 75th-percentile teacher would score about 0.19 standard deviations higher than if he had instead been assigned to the classroom of a 25th-percentile teacher.[28] Thus, the expected student achievement "cost" of one evaluator is approximately 0.19 multiplied by the number of students she would have been teaching instead of serving as an evaluator. While substantial, these costs nevertheless compare favorably to the estimated returns of about 0.11 multiplied by the number of students taught by the several teachers who benefit from the evaluation conducted by that peer evaluator.

VI. Conclusion

The estimates presented here provide evidence that subjective evaluation can spur growth in human capital that improves employee performance even after the evaluation period ends. This is particularly encouraging for the sector we study. In recent years, the consensus among policymakers and researchers has been that after the first few years on the job teacher productivity, at least as measured by student test score growth, cannot be improved. In contrast, we demonstrate that, at least in this setting, experienced teachers provided with relatively detailed information on their performance improved substantially.

American public schools have been under new pressure from regulators and constituents to improve teacher performance. The discussion has focused primarily on evaluation systems as sorting mechanisms: a way to identify the lowest-performing teaches for selective termination. Our work suggests optimism that, while costly, well-structured evaluation systems can not only serve this sorting purpose, but can also improve educational production through nontransient increases in teacher effectiveness. In the language of the education sector, if done well, performance evaluation can be an effective form of teacher professional development.

Finally, to the extent that our results translate to workers in other sectors, public and private, they suggest employers and researchers should reconsider the expected returns to evaluation. Well-designed, well-executed subjective evaluation may affect employee performance through mechanisms other than the proximate, explicit incentives for workers to exert more effort which are the focus of traditional models.

[27] From the perspective of the school district, the replacement is always a new hire. While a principal may be able to replace a peer evaluator with a veteran who transfers from elsewhere in the district, the district will need to replace that transfer with a new hire or let class size grow.

[28] We estimate the standard deviation in overall teacher effect in Cincinnati at 0.14 student-level standard deviations in math. This variation is consistent with estimates from other districts (Hanushek and Rivkin 2010).

REFERENCES

Armstrong, Michael. 2000. "Performance Management." In *Human Resource Management*, edited by Rob Dransfield, 69–84. Oxford, UK: Heinemann Educational Publishers.

Becker, G. S. 1962. "Investment in Human Capital: A Theoretical Analysis." *Journal of Political Economy* 70 (5): 9–49.

Cantrell, Steven, Jon Fullerton, Thomas J. Kane, and Douglas O. Staiger. 2008. "National Board Certification and Teacher Effectiveness: Evidence from a Random Assignment Experiment." National Bureau of Economic Research Working Paper 14608.

Cincinnati Public Schools. 2010. "Cincinnati Public Schools 2010–2011 General Operating Budget." http://www.cps-k12.org/general/finances/BdgtBk1011/BdgtBk1011.pdf (accessed July 4, 2011, along with similar documents for 2005–2006 through 2009–2010).

Clotfelter, Charles T., Helen F. Ladd, and Jacob Vigdor. 2005. "Who Teaches Whom? Race and the Distribution of Novice Teachers." *Economics of Education Review* 24 (4): 377–92.

Clotfelter, Charles T., Helen F. Ladd, and Jacob Vigdor. 2006. "Teacher-Student Matching and the Assessment of Teacher Effectiveness." *Journal of Human Resources* 41 (4): 778–820.

Danielson, Charlotte. 1996. *Enhancing Professional Practice: A Framework for Teaching*. Alexandria, VA: Association for Supervision & Curriculum Development.

Dixit, Avinash. 2002. "Incentives and Organizations in the Public Sector: An Interpretative Review." *Journal of Human Resources* 37 (4): 696–727.

Engellandt, Axel, and Regina T. Riphahn. 2011. "Evidence on Incentive Effects of Subjective Performance Evaluations." *Industrial and Labor Relations Review* 64 (2): 241–57.

Glazerman, Steven, Susanna Loeb, Dan Goldhaber, Douglas Staiger, Stephen Raudenbush, and Grover Whitehurst. 2010. *Evaluating Teachers: The Important Role of Value-Added*. Washington, DC: The Brookings Institution.

Goldhaber, Dan, and Emily Anthony. 2007. "Can Teacher Quality Be Effectively Assessed? National Board Certification as a Signal of Effective Teaching." *Review of Economics and Statistics* 89 (1): 134–50.

Goldhaber, Dan, and Michael Hansen. 2010. "Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions." National Center for Analysis of Longitudinal Data in Education Research Working Paper 31.

Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. *Identifying Effective Teachers Using Performance on the Job*. Washington, DC: The Brookings Institution.

Grossman, Pam, Susanna Loeb, Julia Cohen, Karen Hammerness, James Wyckoff, Donald Boyd, and Hamilton Lankford. 2010. "Measure for Measure: The Relationship between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores." National Bureau of Economic Research Working Paper 16015.

Hanushek, Eric A. 1986. "The Economics of Schooling: Production and Efficiency in Public Schools." *Journal of Economic Literature* 24 (3): 1141–77.

Hanushek, Eric A. 1997. "Assessing the Effects of School Resources on Student Performance: An Update." *Educational Evaluation and Policy Analysis* 19 (2): 141–64.

Hanushek, Eric A. 2011. "Valuing Teachers: How Much Is a Good Teacher Worth?" *Education Next* 11 (3): 41–5.

Hanushek, Eric A., and Steven G. Rivkin. 2010. "Generalizations about Using Value-Added Measures of Teacher Quality." *American Economic Review* 100 (2): 267–71.

Holtzapple, Elizabeth. 2003. "Criterion-Related Validity Evidence for a Standards-Based Teacher Evaluation System." *Journal of Personnel Evaluation in Education* 17 (3): 207–19.

Jackson, C. Kirabo, and Elias Bruegmann. 2009. "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers." *American Economic Journal: Applied Economics* 1 (4): 85–108.

Jacob, Brian A. 2007. "The Challenges of Staffing Urban Schools with Effective Teachers." *Future of Children* 17 (1): 129–53.

Jacob, Brian A., and Lars Lefgren. 2008. "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education." *Journal of Labor Economics* 26 (1): 101–36.

Kane, Thomas J., and Steven Cantrell. 2010. *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project*. Seattle, WA: Bill and Melinda Gates Foundation.

Kane, Thomas J., Eric S. Taylor, John H. Tyler, and Amy L. Wooten. 2011. "Identifying Effective Classroom Practice Using Student Achievement Data." *Journal of Human Resources* 46 (3): 587–613.

Kimball, Steven M. 2002. "Analysis of Feedback, Enabling Conditions and Fairness Perceptions of Teachers in Three School Districts with New Standards-Based Evaluation Systems." *Journal of Personnel Evaluation in Education* 16 (4): 241–68.

Kluger, Avraham N., and Angelo DeNisi. 1996. "The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory." *Psychological Bulletin* 119 (2): 254–84.

Milanowski, Anthony. 2004. "The Relationship between Teacher Performance Evaluation Scores and Student Achievement: Evidence from Cincinnati." *Peabody Journal of Education* 79 (4): 33–53.

Milanowski, Anthony T., and Herbert G. Heneman. 2001. "Assessment of Teacher Reactions to a Standards-Based Teacher Evaluation System: A Pilot Study." *Journal of Personnel Evaluation in Education* 15 (3): 193–212.

Milanowski, Anthony T., Steven M. Kimball, and Brad White. 2004. "The Relationship between Standards-Based Teacher Evaluation Scores and Student Achievement: Replication and Extensions at Three Sites." University of Wisconsin Consortium for Policy Research in Education Working Paper TC-04-01.

Neal, Derek. 2011. "The Design of Performance Pay in Education." In *Handbook of the Economics of Education Volume 4*, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 495–550. Amsterdam: North-Holland, Elsevier.

Papay, John P., and Matthew A. Kraft. 2010. "Do Teachers Continue to Improve with Experience? Evidence of Long-Term Career Growth in the Teacher Labor Market." Paper presented at the Annual Fall Meeting of the Association for Public Policy Analysis and Management, Boston, November 4–6.

Prendergast, Canice. 1999. "The Provision of Incentives in Firms." *Journal of Economic Literature* 37 (1): 7–63.

Rockoff, Jonah E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review* 94 (2): 247–52.

Rockoff, Jonah E., and Cecilia Speroni. 2010. "Subjective and Objective Evaluations of Teacher Effectiveness." *American Economic Review* 100 (2): 261–66.

Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger. 2011. "Can You Recognize an Effective Teacher When You Recruit One?" *Education Finance and Policy* 6 (1): 43–74.

Rockoff, Jonah E., Douglas O. Staiger, Thomas J. Kane, and Eric S. Taylor. 2012. "Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools." *American Economic Review* 102 (7): 3184–3213.

Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics* 125 (1): 175–214.

Springer, Matthew G., Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J. R. Lockwood, Daniel F. McCaffrey, Matthew Pepper, and Brian M. Stecher. 2010. *Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching*. Nashville: National Center on Performance Incentives at Vanderbilt University.

Staiger, Douglas O., and Jonah E. Rockoff. 2010. "Searching for Effective Teachers with Imperfect Information." *Journal of Economic Perspectives* 24 (3): 97–118.

Taylor, Eric S., and John H. Tyler. 2012. "The Effect of Evaluation on Teacher Performance: Dataset." *American Economic Review*. http://dx.doi.org/10.1257/aer.102.7.3628.

Todd, Petra E., and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *Economic Journal* 113 (485): F3–33.

Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, and David Keeling. 2009. *The Widget Effect: Our National Failure to Acknowledge and Act on Teacher Effectiveness*. New York: The New Teacher Project.

Wiswall, Matthew. 2011. "The Dynamics of Teacher Quality." Unpublished.

Yoon, Kwang Suk, Teresa Duncan, Silvia Wen-Yu Lee, Beth Scarloss, and Kathy L. Shapley. 2007. *Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement: Issues & Answers Report, REL 2007–No. 033*. Washington, DC: US Department of Education, Institute of Education Sciences.